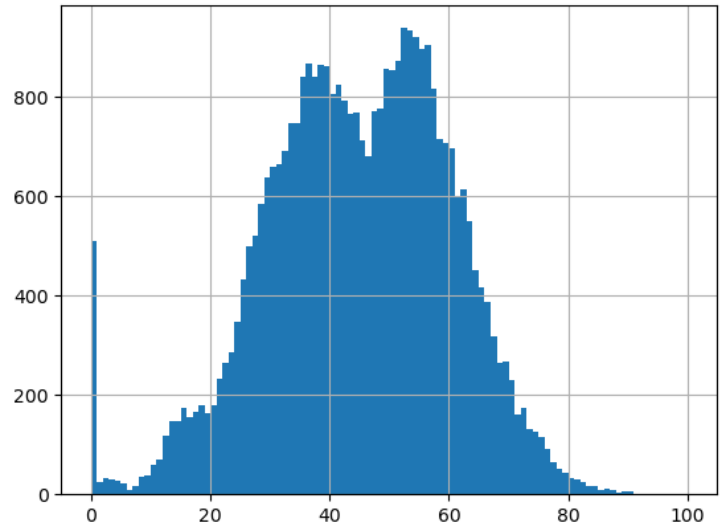
Music genre prediction

Описание:

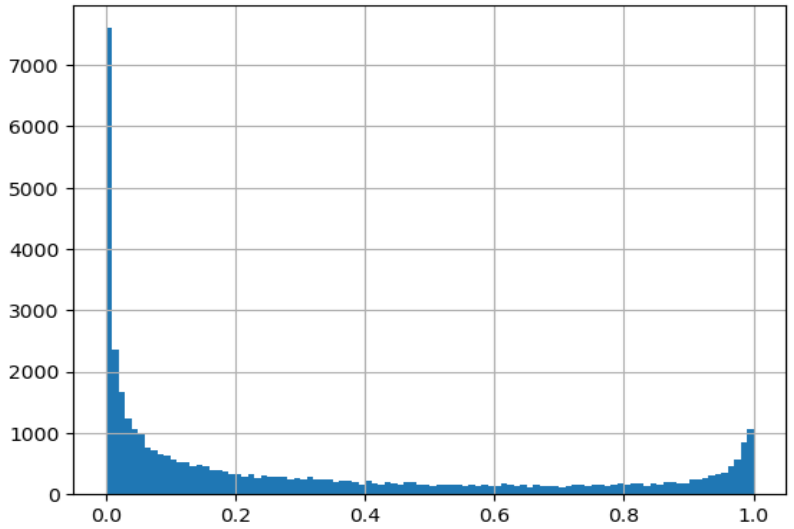
- Задача ставится музыкальным стриминговым сервисом "МиФаСоль". Сервис расширяет работу с новыми артистами и музыкантами, в связи с чем возникла задача - правильно классифицировать новые музыкальные треки, чтобы улучшить работу рекомендательной системы. Был подготовлен датасет, в котором собраны некоторые характеристики музыкальных произведений и их жанры. **Задача - разработать модель, позволяющую классифицировать музыкальные произведения по жанрам.**

Описание полей данных:

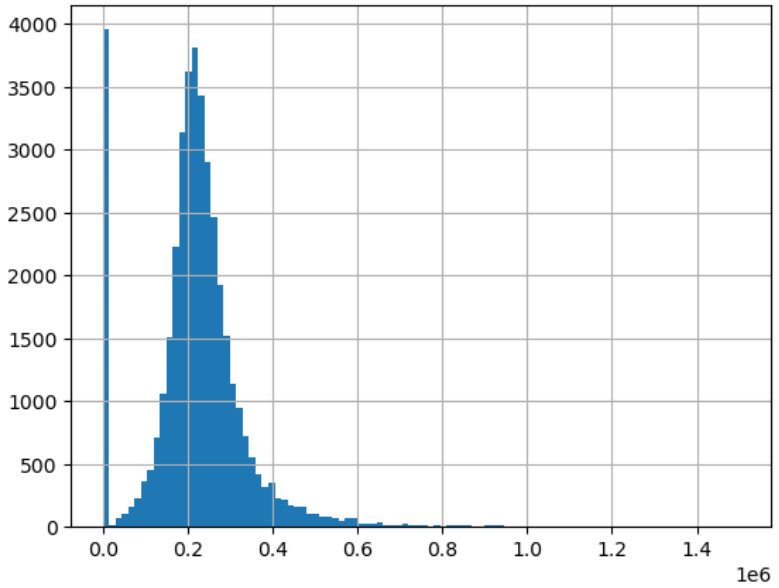
- instance_id -Уникальный идентификатор трека
- track_name - Название трека
- popularity - Популярность трека
- acousticness - Мера уверенности от 0,0 до 1,0 в том, что трек является акустическим. 1,0 означает высокую степень уверенности в том, что трек является акустическим.
- danceability - Танцевальность описывает, насколько трек подходит для танцев, основываясь на сочетании музыкальных элементов, включая темп, стабильность ритма, силу ударов и общую регулярность. Значение 0,0 означает наименьшую танцевальность, а 1,0 - наибольшую танцевальность.
- duration_ms - Продолжительность трека в миллисекундах.
- energy - Энергия это показатель от 0,0 до 1,0, представляющий собой меру интенсивности и активности. Как правило, энергичные композиции ощущаются как быстрые, громкие и шумные. Например, дэт-метал обладает высокой энергией, в то время как прелюдия Баха имеет низкую оценку этого параметра
- instrumentalness - Определяет, содержит ли трек вокал. Звуки "Ooh" и "aah" в данном контексте рассматриваются как инструментальные. Рэп или разговорные треки явно являются "вокальными". Чем ближе значение инструментальности к 1,0, тем больше вероятность того, что трек не содержит вокала
- key - базовый ключ (нота) произведения
- liveness - Определяет присутствие аудитории в записи. Более высокие значения liveness означают увеличение вероятности того, что трек был исполнен вживую. Значение выше 0,8 обеспечивает высокую вероятность того, что трек исполняется вживую
- loudness - Общая громкость трека в децибелах (дБ)
- mode - Указывает на модальность (мажорную или минорную) трека
- speechiness - Речевой характер определяет наличие в треке разговорной речи. Чем более исключительно речевой характер носит запись (например, ток-шоу, аудиокнига, поэзия), тем ближе значение атрибута к 1,0. Значения выше 0,66 характеризуют треки, которые, вероятно, полностью состоят из разговорной речи. Значения от 0,33 до 0,66 характеризуют треки, которые могут содержать как музыку, так и речь, как в виде фрагментов, так и в виде слоев, включая такие случаи, как рэп-музыка. Значения ниже 0,33, скорее всего, представляют музыку и другие неречевые треки.
- tempo - Темп трека в ударах в минуту (BPM). В музыкальной терминологии темп представляет собой скорость или темп данного произведения и напрямую зависит от средней продолжительности тактов
- obtained_date - дата загрузки в сервис
- valence - Показатель от 0,0 до 1,0, характеризующий музыкальный позитив, передаваемый треком. Композиции с высокой валентностью звучат более позитивно (например, радостно, весело, эйфорично), а композиции с низкой валентностью - более негативно (например, грустно, депрессивно, сердито)
- music_genre - Музыкальный жанр трека



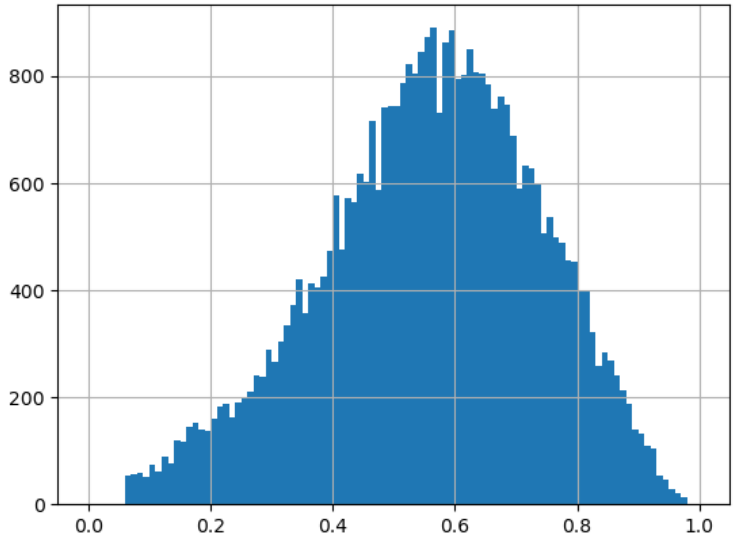
popularity - Популярность трека



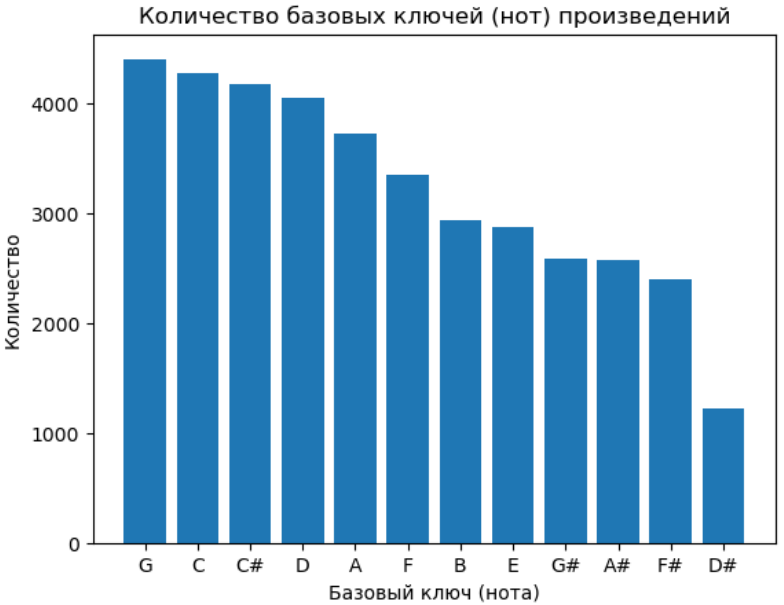
acousticness - Мера уверенности от 0,0 до 1,0 в том, что трек является акустическим. 1,0 означает высокую степень уверенности в том, что трек является акустическим.



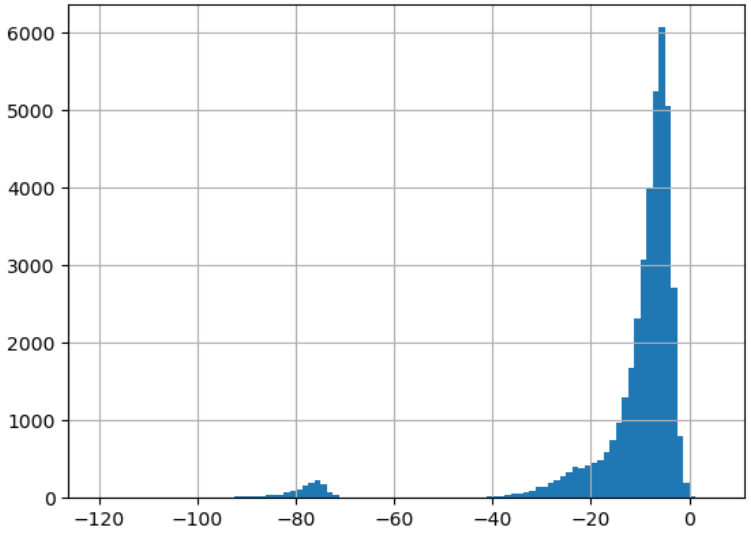
duration_ms - Продолжительность трека в миллисекундах.



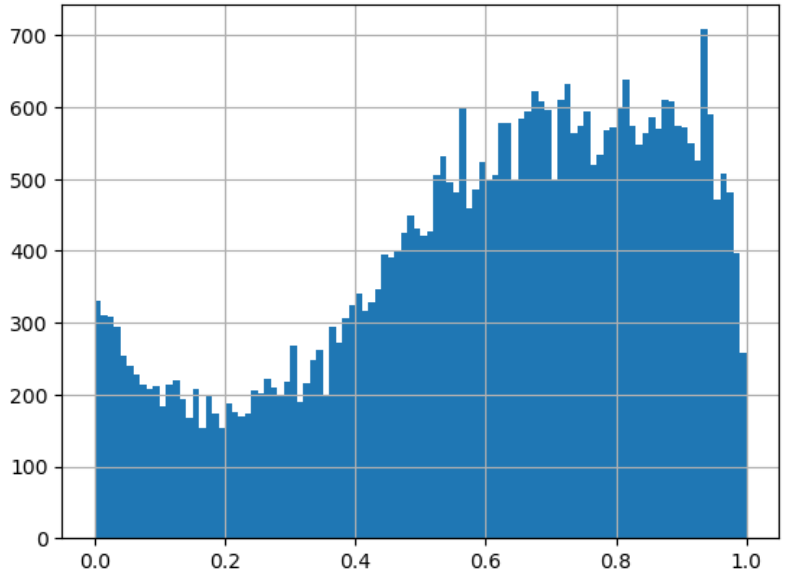
danceability - Танцевальность описывает, насколько трек подходит для танцев, основываясь на сочетании музыкальных элементов, включая темп, стабильность ритма, силу ударов и общую регулярность. Значение 0,0 означает наименьшую танцевальность, а 1,0 - наибольшую танцевальность.



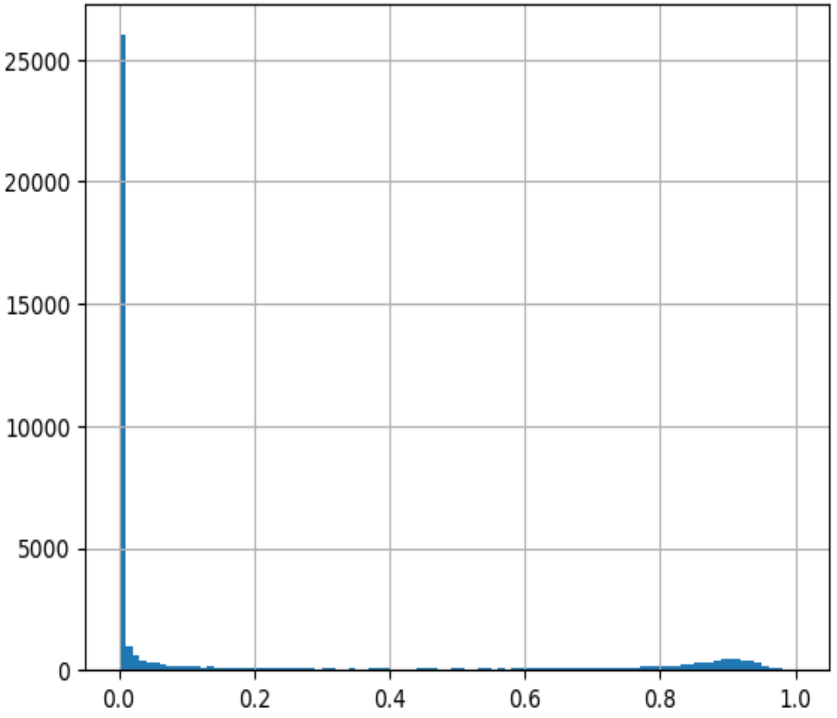
key - базовый ключ (нота) произведения



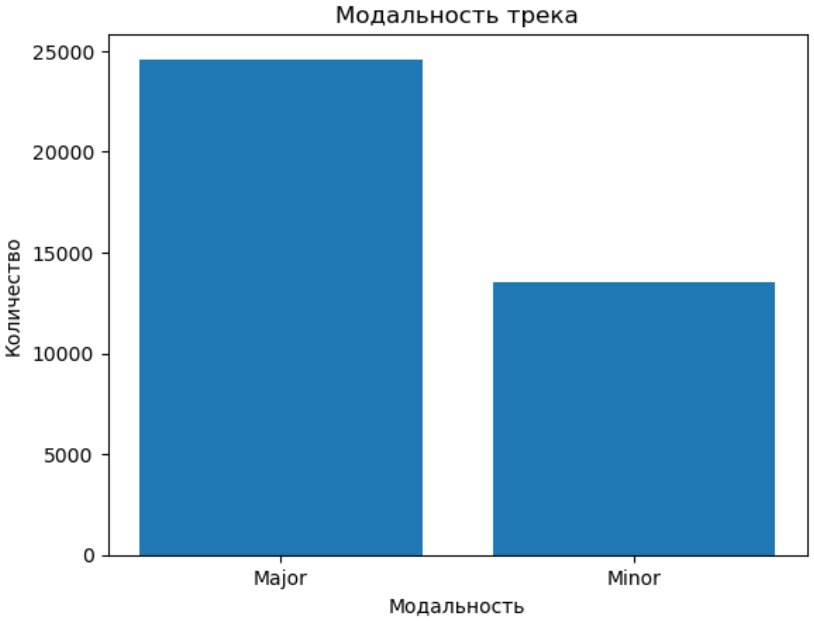
loudness - Общая громкость трека в децибелах (дБ)



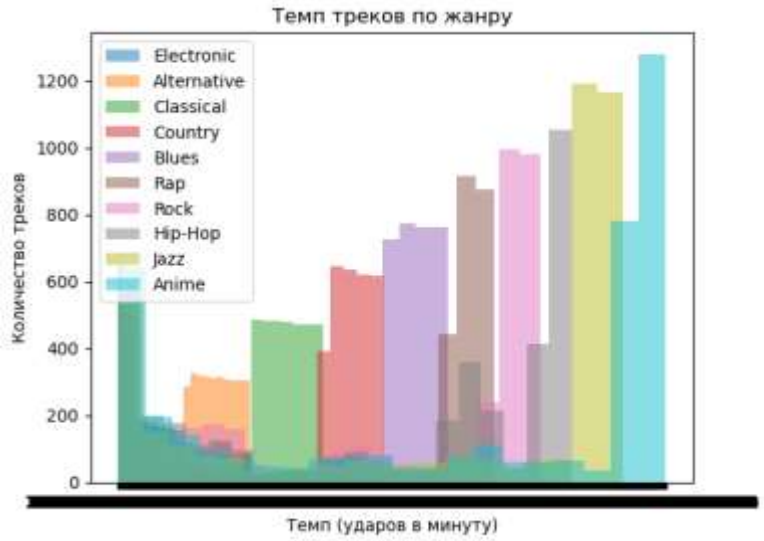
energy - Энергия это показатель от 0,0 до 1,0, представляющий собой меру интенсивности и активности. Как правило, энергичные композиции ощущаются как быстрые, громкие и шумные. Например, дэт-метал обладает высокой энергией, в то время как прелюдия Баха имеет низкую оценку этого параметра



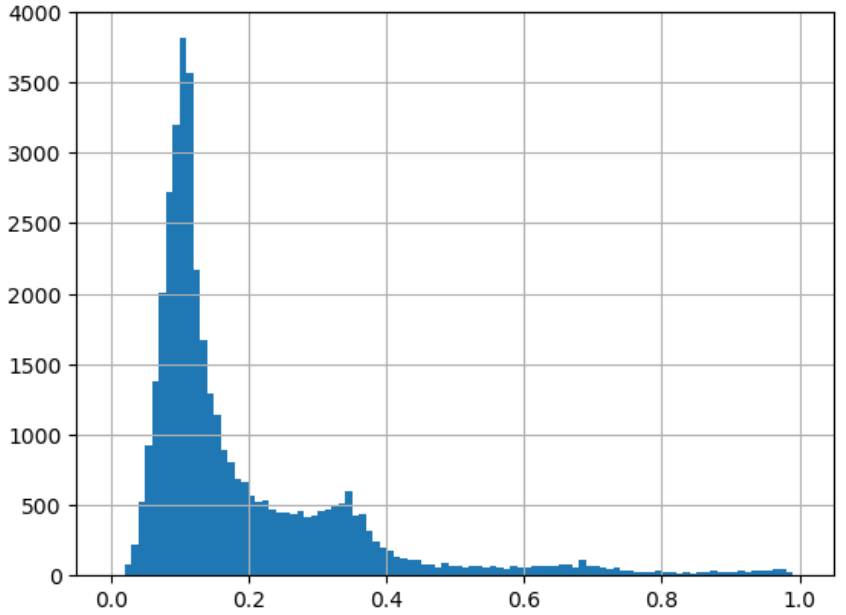
instrumentalness - Определяет, содержит ли трек вокал. Звуки "Ooh" и "aah" в данном контексте рассматриваются как инструментальные. Рэп или разговорные треки явно являются "вокальными". Чем ближе значение инструментальности к 1,0, тем больше вероятность того, что трек не содержит вокала



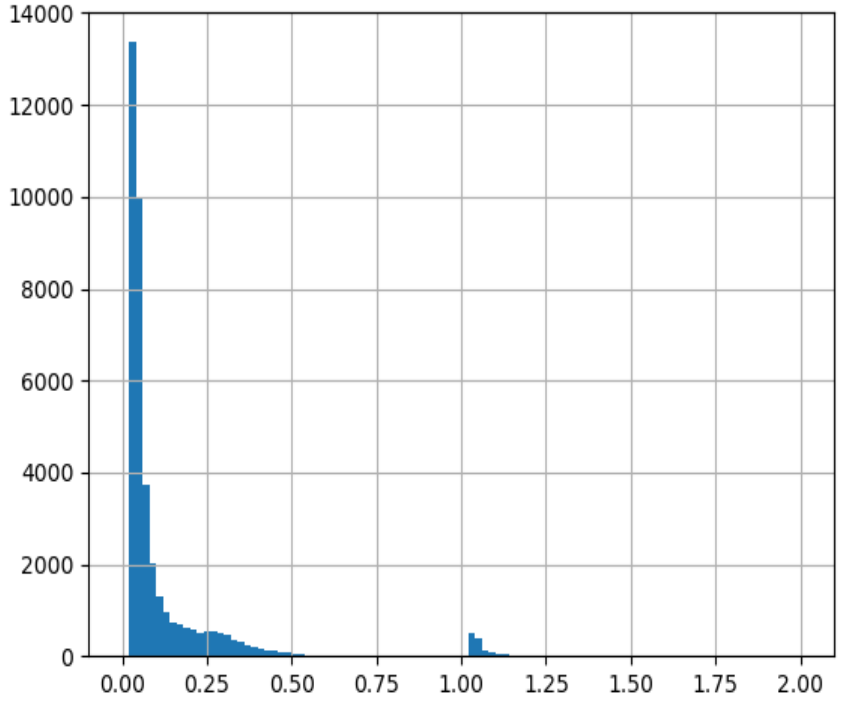
mode - Указывает на модальность (мажорную или минорную) трека



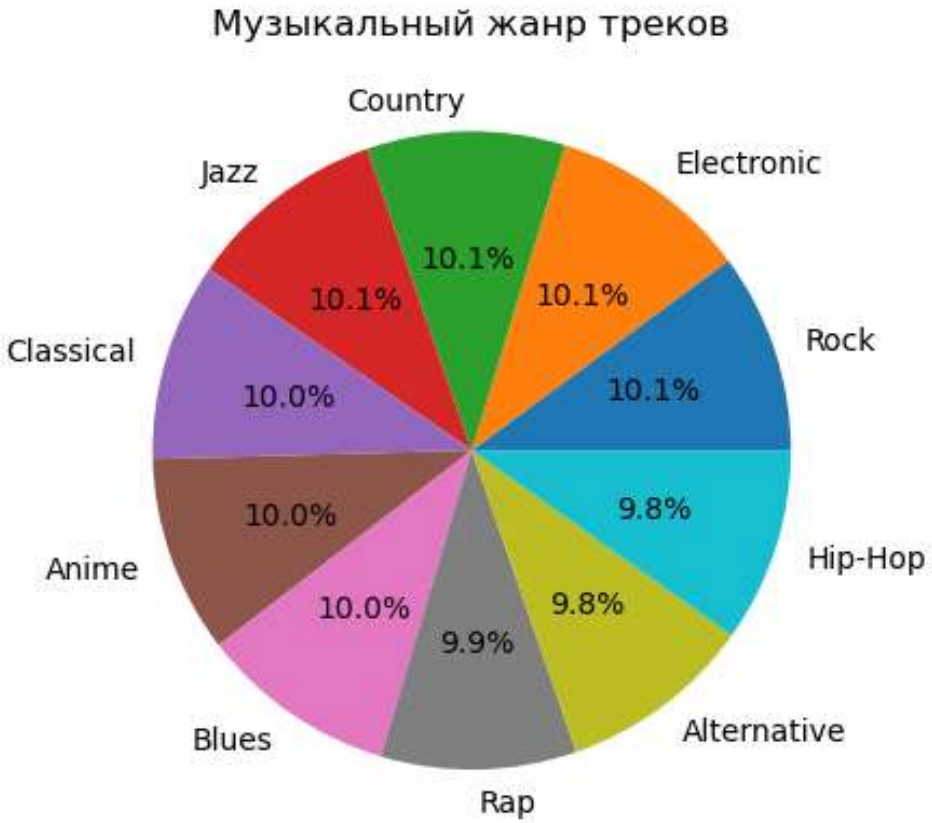
tempo - Темп трека в ударах в минуту (BPM).



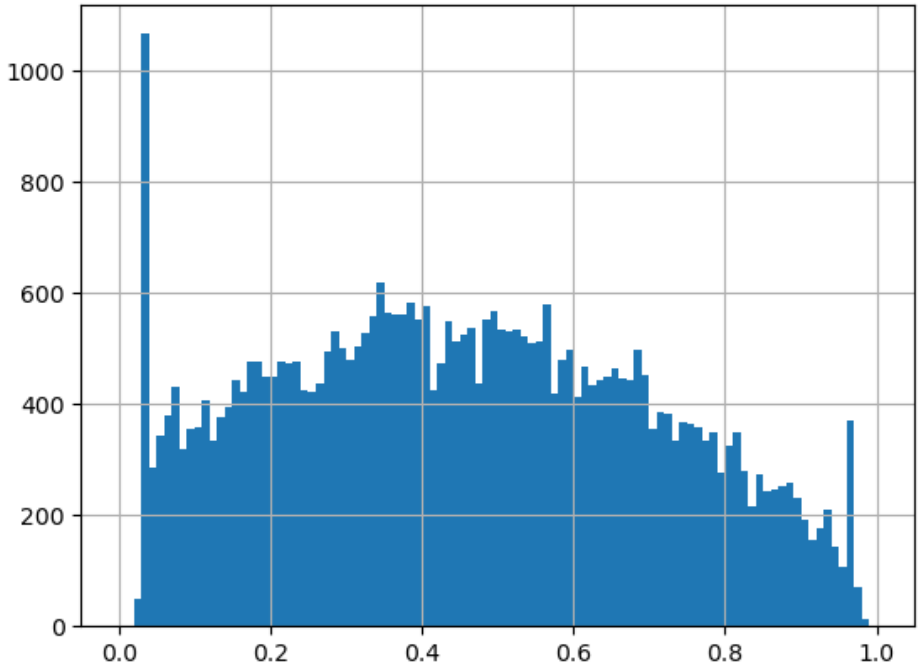
liveness - Определяет присутствие аудитории в записи. Более высокие значения liveness означают увеличение вероятности того, что трек был исполнен вживую. Значение выше 0,8 обеспечивает высокую вероятность того, что трек исполняется вживую



speechiness - Речевой характер определяет наличие в треке разговорной речи. Чем более исключительно речевой характер носит запись (например, ток-шоу, аудиокнига, поэзия), тем ближе значение атрибута к 1,0. Значения выше 0,66 характеризуют треки, которые, вероятно, полностью состоят из разговорной речи. Значения от 0,33 до 0,66 характеризуют треки, которые могут содержать как музыку, так и речь, как в виде фрагментов, так и в виде слоев, включая такие случаи, как рэп-музыка. Значения ниже 0,33, скорее всего, представляют музыку и другие неречевые треки.



***music_genre* - Музыкальный жанр трека**



***valence* - Показатель от 0,0 до 1,0, характеризующий музыкальный позитив, передаваемый треком. Композиции с высокой валентностью звучат более позитивно (например, радостно, весело, эйфорично), а композиции с низкой валентностью - более негативно (например, грустно, депрессивно, сердито)**

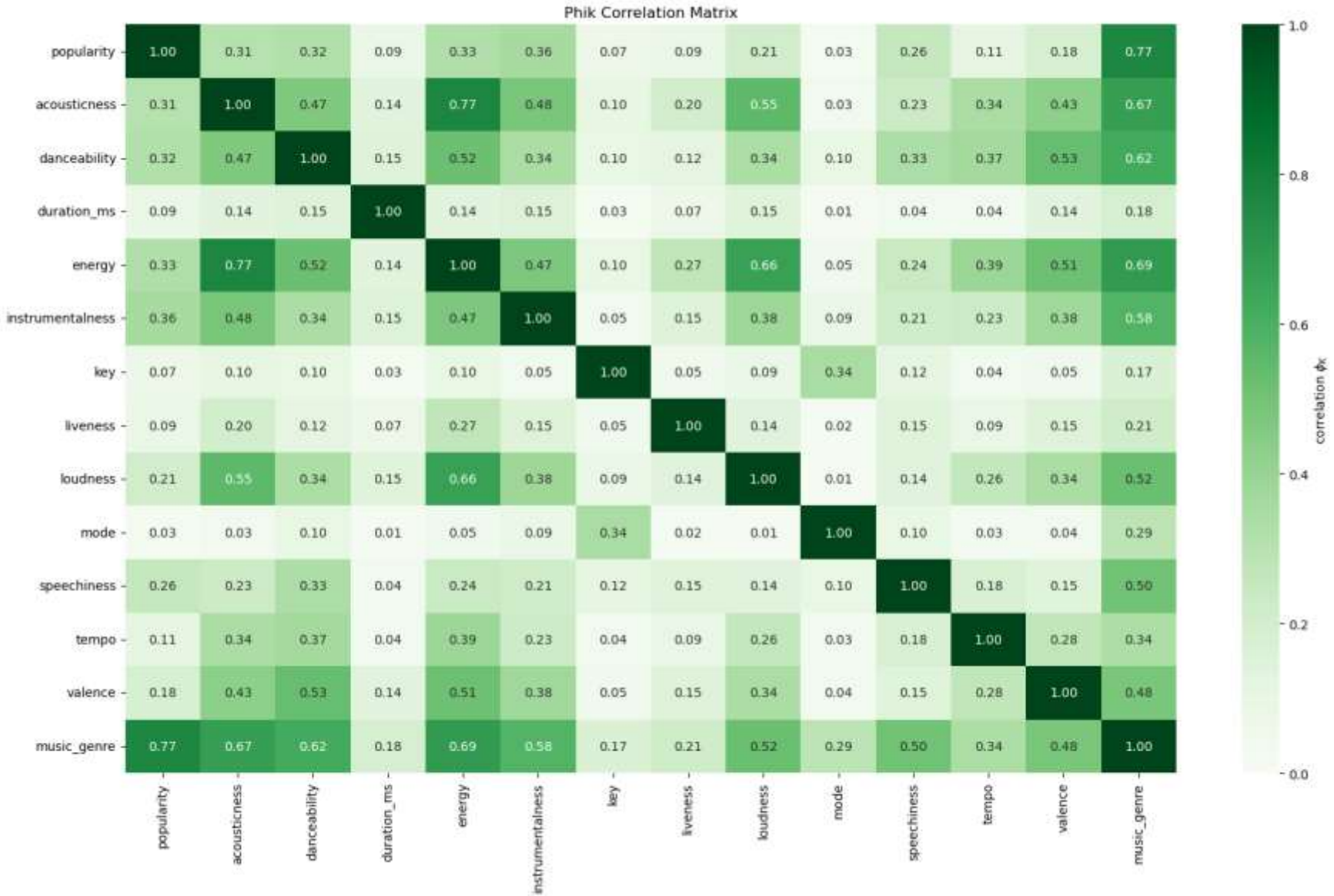
Обработка данных и полноценный разведочный анализ:

- Проверили на явные дубликаты, их не нашлось, там где представилось возможным - заменили пропуски, обработали аномальные значения.

После проведения полноценного разведочного анализа смогли отметить следующее:

- Самый популярный трек оказался безымянным
- Популярность трека в большинстве своем составляет в районе 40%
- Большая часть треков имеет меру уверенности в том, что трек является акустическим, равную нулю
- Целых 69 треков почти абсолютно точно являются аккустическими
- Самый танцевальный трек - Cherish My Dawgs в жанре Hip-Hop
- В среднем мера танцевальности трека составляет 0.53
- Самая длительная композиция на целых 1.3 часа - это Lost Lands 2017 Mix в жанре Electronic
- Самые неэнергичные композиции в жанре - Classical
- Самые энергичные композиции в жанрах - Anime и Electronic
- Самый популярный вокальный жанр - Country
- Самый масштабный невокальный жанр - Classical
- Вживую чаще всего исполняются композиции в жанре - Blues
- Большая часть треков имеют модальность - Major
- Треки, которые, вероятно, полностью состоят из разговорной речи в большинстве своем в жанре - Hip-Hop
- Значения speechiness от 0,33 до 0,66 характеризуют треки, которые могут содержать как музыку, так и речь, больше всего представителей Hip-Hop и Rap жанров
- Значения speechiness ниже 0,33, скорее всего, представляют музыку и другие неречевые треки - тут у нас целое изобилие музыкальных жанров: Rock, Country,Classical, Anime, Blues, Jazz, Electronic, Alternative
- Самые грустные композиции в жанре - Classical
- Самый позитивный и веселый жанр - Blues
- У нас 10 музыкальных жанров, преобладание какого-то конкретного не наблюдается, количество каждого жанра в процентном соотношении почти одинаково

Проверка на мультиколлинеарность



- Из вывода видно, что существует несколько пар признаков с высокой мультиколлинеарностью, определенной с использованием корреляции по методу phik. Каждая строка в выводе представляет собой пару признаков, которые имеют высокую корреляцию между собой.
- Например, если взять первую строку "popularity and music_genre have high phik correlation.", это означает, что признаки popularity и music_genre имеют высокую корреляцию по методу phik, дальше уделяли внимание признакам с высокой корреляцией, так как было опасение, что они смогут вносить смещения в модели.

Выбор и обучение моделей

```
Grid Search Progress:   0%|                                        | 0/3 [00:00<?, ?it/s]
Searching for RandomForest...
Grid Search Progress:  33%|[redacted]                | 1/3 [03:56<07:52, 236.31s/it]
Best parameters for RandomForest: {'classifier__n_estimators': 100, 'classifier__min_samples_split': 10, 'classifie
r__min_samples_leaf': 4, 'classifier__max_depth': 15}
Time elapsed for RandomForest: 236.16368508338928 seconds
F1 score for RandomForest: 0.54225

Searching for SVM...
Best parameters for SVM: {'classifier__kernel': 'rbf', 'classifier__gamma': 0.1, 'classifier__C': 1}
Time elapsed for SVM: 1644.3382592201233 seconds
Grid Search Progress:  67%|[redacted]        | 2/3 [31:30<17:50, 1070.35s/it]
F1 score for SVM: 0.538625

Searching for LogisticRegression...
The total space of parameters 6 is smaller than n_iter=10. Running 6 iterations. For exhaustive searches, use GridS
earchCV.
Grid Search Progress: 100%|[redacted]| 3/3 [31:46<00:00, 635.60s/it]
Best parameters for LogisticRegression: {'classifier__solver': 'liblinear', 'classifier__penalty': 'l2', 'classifie
r__max_iter': 100, 'classifier__C': 0.1}
Time elapsed for LogisticRegression: 16.242199182510376 seconds
F1 score for LogisticRegression: 0.421125
```

- Создали pipeline для обработки признаков и обучения трех моделей: RandomForest, SVM и LogisticRegression. В коде использовался RandomizedSearchCV для поиска лучших параметров для каждой модели и вывода результаты. Метрика f1_micro использовалась для оценки качества моделей.
- Выполнили преобразования для числовых и категориальных признаков.Выполняли замену пропущенных значений средним и стандартизацию числовых признаков, а для категориальных заменяли пропущенные значения наиболее часто встречающимся и применяли one-hot encoding. С помощью ColumnTransformer объединили эти преобразования.

Значение метрики F1_micro и лучшие параметры для моделей:

- RandomForest: F1_micro: 0.54225
- Best parameters for RandomForest: {'classifier__n_estimators': 100, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 4, 'classifier__max_depth': 15}
- SVM: F1_micro: 0.538625
- Best parameters for SVM: {'classifier__kernel': 'rbf', 'classifier__gamma': 0.1, 'classifier__C': 1}
- LogisticRegression: F1_micro: 0.421125
- Best parameters for LogisticRegression: {'classifier__solver': 'liblinear', 'classifier__penalty': 'l2', 'classifier__max_iter': 100, 'classifier__C': 0.1}

Лучшее значение метрики F1_micro можно сказать, что достигнуто моделями SVM и RandomForest, поэтому выбрали их для финальной оценки.

Итоговые оценки и выбор лучшей модели:

- На Kaggle модель SVM дала значение в 0.5228 с лучшими параметрами {'classifier__kernel': 'rbf', 'classifier__gamma': 0.01, 'classifier__C': 10}, а RandomForest с лучшими параметрами {'classifier__n_estimators': 100, 'classifier__min_samples_split': 10, 'classifier__min_samples_leaf': 4, 'classifier__max_depth': 15} - 0.5436, поэтому было принято решение дальше поперебирать модели с целью улучшения значения.
- Получили, что F1_micro: 0.57925 с лучшими параметрами для CatBoost: {'classifier__learning_rate': 0.1, 'classifier__iterations': 300, 'classifier__depth': 6}
- Получили, что F1_micro: 0.571125 с лучшими параметрами для XGBoost: {'classifier__n_estimators': 200, 'classifier__max_depth': 4, 'classifier__learning_rate': 0.1}
- Получили, что F1_micro: 0.569875 с лучшими параметрами для LightGBM: {'classifier__n_estimators': 300, 'classifier__max_depth': 4, 'classifier__learning_rate': 0.05}

Для модели CatBoost с лучшими параметрами на Kaggle составила 0.5732, из всех опробованных моделей это лучший результат. Так что оставили эту модель с ее лучшими параметрами как финальную.

Анализ важности признаков